INFORMATION AND CONTROL 8, 264–281 (1965)

An Asymptotic Simultaneous Diagonalization Procedure for Pattern Recognition

JACK CAPON

Applied Research Laboratory, Sylvania Electronic Systems, Waltham, Massachusetts

The problem of classifying an observation into one of several different covariance matrices, of Toeplitz type, is considered. These matrices occur when the observation consists of uniformly spaced samples from a weakly stationary stochastic process. The usual Bayes procedure for this problem, assuming Gaussian distributions and a zero-one loss function, partitions the sample space with quadratic forms and is very difficult to implement.

An asymptotic simultaneous diagonalization procedure (ASDP) is introduced which, in an asymptotic sense, simultaneously diagonalizes all of the covariance matrices. This result is valid regardless of whether the Gaussian assumption is made. A classification method based on the ASDP is proposed which can be used regardless of whether the Gaussian assumption holds. In the Gaussian case it is shown, using the theory of Toeplitz forms, that the performance of the ASDP classification system compares very favorably, in an asymptotic sense, with that of the optimum, and more complicated, Bayes procedure.

The ASDP classification system has considerable appeal from an engineering standpoint. The system essentially estimates the sampled power spectrum by applying the observation to a bank of orthogonal, and uniformly spaced, narrow-band filters. This estimate is weighted by values of sampled power spectra and combined with bias terms in order to obtain a classification decision. A comparison of the present results with some related work of Price and Kailath is also given.

I. INTRODUCTION

The pattern recognition problem is of considerable importance in engineering applications. Many examples can be cited, such as speech and character recognition, and target identification. It has been recognized that the pattern recognition problem can be discussed within the framework of statistical classification theory. Thus, let us suppose that there exists a set of patterns, or categories, which we desire to recognize by artificial or mechanical means. We denote these categories by $C_1, \cdots, C_m$. In addition, we assume that we are given a set of observations, or measurements, $X_1, \cdots, X_N$, which can be considered to be a set of N continuous, real, random variables. The pattern recognition, or statistical classification, problem consists of partitioning the measurement space into m regions $R_1, \cdots, R_m$, so that when $X_1, \cdots, X_N$ lies in R_k we decide that the measurements came from the category, or pattern, C_k.

The reliability of a particular recognition procedure can be described quantitatively by means of certain misclassification probabilities and cost functions. These in turn define a loss function, which usually forms the basis for choosing the classification regions. For example, the Bayes procedure chooses $R_1, \cdots, R_m$ so as to minimize the expected loss, while the minimax procedure minimizes the maximum expected loss, cf. Anderson (1958), chap. 6.

Much of the literature on statistical classification theory is based on the assumption that within each class C_k there is associated a multivariate Gaussian probability density for the measurements. This assumption is necessitated by the fact that it is difficult to write the multivariate probability density if it is not Gaussian. It is well known that under this assumption the Bayes, as well as the minimax, procedure leads to a partitioning of the measurement space by boundaries which are quadratic functions, assuming, of course, a zero-one loss function. This method of partitioning the measurement space is difficult to implement, in general, particularly if N happens to be large. While the scheme leans very heavily on the assumption that the multivariate distribution of the measurements is Gaussian, for each C_k, it is still possible to justify its use when this assumption is dropped.

Another procedure has been considered by Anderson and Bahadur (1962) which is independent of the Gaussian assumption and results in a partitioning of the measurement space with linear functions or plane boundaries. This method of partitioning is simpler to implement than the previous one which used quadratic boundaries. However, the reliability of the Anderson-Bahadur procedure must always be less, or at best the same, as the Bayes or minimax procedures. An application of both the Bayes and Anderson-Bahadur procedures to the vowel recognition problem has been made by Welch and Wimpress (1961).

In the present work we will consider the pattern recognition problem when the covariance matrix of the measurements is a Toeplitz matrix, for each category C_k. These matrices arise whenever the observations are uniformly spaced samples from a weakly stationary stochastic process, which includes enough applications to be of practical significance. The optimum Bayes procedure for this case is still quite difficult to implement, since the partitioning of the measurement space is still done with quadratic functions. However, a simple linear unitary transformation of the measurements will be introduced which, in an asymptotic sense, simultaneously diagonalizes all of the covariance matrices, and allows the partitioning of the measurement space to be implemented quite easily. This result is independent of the Gaussian assumption so that the procedure is applicable for both the non-Gaussian as well as Gaussian cases. In the Gaussian case it is shown that the reliability of this procedure compares favorably, in an asymptotic sense, with the optimum and more complicated Bayes procedure.

II. OPTIMUM CLASSIFICATION PROCEDURES

We consider the problem of classifying an observation,

$$\mathbf{x}_N = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix},$$

into one of several categories, or patterns, $C_1, \cdots, C_m$. In addition, we assume that each category C_k has associated with it a probability density function $p_k(\mathbf{x}_N)$. It is desired to divide the space of observations into m mutually exclusive and exhaustive regions $R_1, \cdots, R_m$. If an observation falls into R_k we shall say that it comes from C_k.

Let the cost of misclassifying an observation from C_k as coming from C_j be $C(j \mid k)$. The probability of this misclassification is

$$P(j \mid k, R) = \int_{R_j} p_k(\mathbf{x}_N)\, d\mathbf{x}_N. \tag{1}$$

The probability of misclassifying an observation from C_k is

$$P'(k, R) = \sum_{\substack{j=1 \\ j \neq k}}^{m} P(j \mid k, R). \tag{2}$$

If $q_1, \cdots, q_m$ denote the a priori probabilities of the categories, then the expected loss is

$$\sum_{k=1}^{m} q_k \left\{ \sum_{\substack{j=1 \\ j \neq k}}^{m} C(j \mid k) P(j \mid k, R) \right\}. \tag{3}$$

In order to simplify matters only the Bayes procedure will be considered. The Bayes procedure chooses $R_1, \cdots, R_m$ so as to minimize the expected loss given in (3). If we assume that

$$\begin{aligned} C(j \mid k) &= 1, \qquad j, k = 1, \cdots, m, j \neq k, \\ &= 0, \qquad j = k, j = 1, \cdots, m, \end{aligned} \tag{4}$$

and all patterns are a priori equally probable so that

$$q_k = 1/m, \qquad k = 1, \cdots, m, \tag{5}$$

then the Bayes procedure assigns $\mathbf{x}_N$ to R_k if

$$p_j(\mathbf{x}_N) < p_k(\mathbf{x}_N), \qquad j = 1, \cdots, m, j \neq k, \tag{6}$$

or

$$Q_{jk} = \ln \frac{p_k(\mathbf{x}_N)}{p_j(\mathbf{x}_N)} > 0, \qquad j = 1, \cdots, m, j \neq k, \tag{7}$$

cf., e.g., Anderson (1958), p. 144. That is, the point $\mathbf{x}_N$ is in C_k if k is the index for which $p_j(\mathbf{x}_N)$ is a maximum, or in other words C_k is the most probable population.

We will assume that we have multivariate Gaussian probability densities

$$p_k(\mathbf{x}_N) = (2\pi)^{-N/2} \mid \Sigma_{Nk} \mid^{-1/2} \exp\left(-\tfrac{1}{2}\mathbf{x}_N{}'\Sigma_{Nk}^{-1}\mathbf{x}_N\right), \qquad k = 1, \cdots, m, \tag{8}$$

where Σ_{Nk} is the covariance matrix of $\mathbf{x}_N$ and is assumed to be positive definite, $\mathbf{x}_N{}'$ denotes transpose, and $\mid \Sigma_{Nk} \mid$ denotes determinant of Σ_{Nk}. It should be noted that there is no essential loss of generality in assuming that the mean value matrix of $\mathbf{x}_N$ is zero, under C_k, $k = 1, \cdots, m$. Thus, we have

$$Q_{jk} = \tfrac{1}{2}\left[\ln \frac{\mid \Sigma_{Nj} \mid}{\mid \Sigma_{Nk} \mid} + \mathbf{x}_N{}'\left(\Sigma_{Nj}^{-1} - \Sigma_{Nk}^{-1}\right)\mathbf{x}_N\right], \quad j, k = 1 \cdots, m, \tag{9}$$

and the best set of regions of classification are $R_1 : Q_{21} > 0, \cdots, Q_{m1} > 0$; $R_2 : Q_{12} > 0, Q_{32} > 0, \cdots, Q_{m2} > 0; \cdots ; R_m : Q_{1m} > 0, \cdots, Q_{m-1,m} > 0$.

The Bayes procedure was derived under the assumption that $\mathbf{x}_N$ has a multivariate Gaussian distribution in C_k, $k = 1, \cdots, m$. The procedure is still a reasonable one to use even if this assumption is not satisfied, provided, of course, that the covariance matrix is still given by Σ_{Nk}, $k = 1, \cdots, m$. However, it is quite complicated to implement the test since it involves the computation of inverses and determinants of m

high-dimensional matrices, the evaluation of m quadratic forms, and the partitioning of the measurement space with quadratic surfaces.

It is well known that there exists a unitary matrix which diagonalizes a given real symmetric matrix. If we are given two real symmetric matrices it is possible to simultaneously diagonalize both matrices with a single unitary matrix if and only if the matrices commute, cf. Perlis (1952), p. 213. The commutation of two matrices is certainly a fortunate occurrence, so that, a fortiori, if we have m real symmetric matrices it would be extremely fortuitous if we could find a unitary matrix which simultaneously diagonalized all m matrices. However, if such a simultaneous diagonalization were possible for $\Sigma_{N1}, \cdots, \Sigma_{Nm}$, the implementation of the Bayes procedure would be considerably simplified, since all quadratic forms could be reduced to sums of squares in terms of the transformed variates, and the problem of finding determinants and inverses would be replaced by the simpler problem of computing eigenvalues. The problem of determining eigenvalues is, of course, simple when such a simultaneous diagonalization procedure exists and may not be simple otherwise.

In the present work we will introduce a unitary matrix which will simultaneously diagonalize $\Sigma_{N1}, \cdots, \Sigma_{Nm}$, in an asymptotic sense which will be made more precise subsequently. The results will be valid only when $\Sigma_{N1}, \cdots, \Sigma_{Nm}$ are Toeplitz matrices. These matrices occur whenever the X_j's are uniformly spaced samples from a weakly stationary stochastic process. This includes enough problems to be of practical significance.

III. THEORY OF TOEPLITZ FORMS

Let $\sigma_{jj'}^{(k)} = E_k(X_j X_{j'}), j, j' = 1, \cdots, N, k = 1, \cdots, m$, where E_k denotes expectation taken with respect to $p_k(\mathbf{x}_N)$. A matrix

$$\Sigma_{Nk} = \{\sigma_{jj'}^{(k)}\}, \qquad j, j' = 1, \cdots, N, \quad k = 1, \cdots, m, \tag{10}$$

will be said to be a Toeplitz matrix if

$$\sigma_{hh'} = \sigma_{jj'}, \quad \text{when} \quad |h - h'| = |j - j'|, \qquad h, h', j, j' = 1, \cdots, N. \tag{11}$$

Let $\rho_{j-j'}^{(k)} = \sigma_{jj'}^{(k)}$, and let $f_k(x)$ be a real-valued function of the class L, i.e., $L_1(-\pi, \pi)$ and

$$f_k(x) \sim \sum_{j=-\infty}^{\infty} \rho_j^{(k)} \epsilon^{ijx}, \qquad i = \sqrt{-1}, \tag{12}$$

its Fourier series, where

$$\rho_j^{(k)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f_k(x) \epsilon^{-ijx} \, dx, \qquad j = 0, \pm 1, \pm 2, \cdots, \tag{13}$$

and are assumed to be known for $-\infty < j < \infty$. Since $\rho_j^{(k)} = \rho_{-j}^{(k)}$, we have $f_k(x) = f_k(-x)$.

The Hermitian form

$$\begin{aligned} T_N(f_k) &= \sum_{j,j'=1}^{N} \rho_{j-j'}^{(k)} u_j u_{j'}^* \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{j=1}^{N} u_j \epsilon^{ijx} \right|^2 f_k(x) \, dx, \end{aligned} \tag{14}$$

is defined as the Toeplitz form associated with the function $f_k(x)$, where u^* denotes complex conjugate. If $X_1, \cdots, X_N$ are uniformly spaced samples from a weakly stationary process $\{X(t)\}$, with spacing T, then $f_k(\omega T)$ is the sampled power spectrum of $\{X(t)\}$, cf. Ragazzini and Franklin (1958), p. 257. An extensive account of the theory of Toeplitz forms has been given by Grenander and Szegö (1958).

It is known, cf. Grenander and Szegö (1958), p. 19, that $f_k(x)$ is nonnegative except for a set of Lebesgue measure zero if and only if its Toeplitz forms are nonnegative for all values of N. Since Σ_{Nk} is a covariance matrix, and thus nonnegative definite for all N, it follows that $f_k(x)$ is nonnegative almost everywhere with respect to Lebesgue measure. In addition, if $f_k(x)$ is positive on a set of positive measure it can easily be shown that Σ_{Nk} is positive definite for each N. In fact, suppose $T_N(f_k) = 0$, then $K_N(x) = \sum_{j=1}^{N} u_j \epsilon^{ijx}$ vanishes on a set of positive measure. However, $K_N(x)$ extends to an entire analytic function of the complex variable z, so that $K_N(x) = 0$ everywhere. Due to the orthogonality of ϵ^{ijx}, $j = 1, \cdots, N$, on $[-\pi, \pi]$, it follows that $u_j = 0$, $j = 1, \cdots, N$, so that Σ_{Nk} is positive definite for each N.

In the ensuing discussions we will need the following regularity conditions for $f_k(x)$, $k = 1, \cdots, m$.

A. $f_k(x)$ is continuously differentiable, $x \in [-\pi, \pi]$. This implies $M_k = \max_{-\pi \leq x \leq \pi} f_k(x)$ is bounded, and also allows us to use the mean value theorem to write

$$f_k(x) = f_k(x_0) + (x - x_0) f_k'(\hat{x}), \tag{15}$$

where $\hat{x}$ is between x and x_0. In addition, this condition implies that $f_k(x)$ is Riemann integrable so that we can always write

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} f_k^{\,s}\left(j\,\frac{2\pi}{N}\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f_k^{\,s}(x)\; dx, \tag{16}$$

for any positive integer s, and also

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \ln f_k\left(j\,\frac{2\pi}{N}\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f_k(x)\; dx, \tag{17}$$

assuming condition B is satisfied.

B. $$m_k = \min_{-\pi \leqq x \leqq \pi} f_k(x) > 0.$$

C. The covariance matrix $B_{m-1,k} = \{b_{jj'}^{(k)}\}$ is positive definite, where

$$2b_{jj'}^{(k)} = \int_{-\pi}^{\pi} \left(\frac{f_k(x)}{f_j(x)} - 1\right)\left(\frac{f_k(x)}{f_{j'}(x)} - 1\right) dx \Big/ \left[\int_{-\pi}^{\pi} \left(\frac{f_k(x)}{f_j(x)} - 1\right)^2 dx\right]^{1/2} \left[\int_{-\pi}^{\pi} \left(\frac{f_k(x)}{f_{j'}(x)} - 1\right)^2 dx\right]^{1/2}, \tag{18}$$

$$j, j' = 1, \cdots, m; \qquad j, j' \neq k.$$

D. $|f_k(x) - f_j(x)| > 0$ on a set of positive measure, $x \in [-\pi, \pi]$, all $j, k = 1, \cdots, m, j \neq k$.

Let $\lambda_{N1}^{(k)}, \lambda_{N2}^{(k)}, \cdots, \lambda_{NN}^{(k)}$, denote the eigenvalues of the matrix Σ_{Nk}. If $F(\lambda)$ is any continuous function defined in the finite interval $m_k \leqq \lambda \leqq M_k$, then we have, cf. Grenander and Szegö, (1958), p. 65,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} F(\lambda_{Nj}^{(k)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f_k(x))\; dx. \tag{19}$$

Let $\lambda_{N1}^{(jk)}, \lambda_{N2}^{(jk)}, \cdots, \lambda_{NN}^{(jk)}$, denote the eigenvalues of $\Sigma_{Nj}\Sigma_{Nk}^{-1}$. Then, according to Grenander and Szegö, pp. 105–106, we have, by observing that if $a_1 \leqq \cdots \leqq a_n$, $b_1 \leqq \cdots \leqq b_n$ are the eigenvalues of the positive definite matrices A, B, respectively, and if these matrices commute, i.e., $AB = BA$, then $a_1b_1 \leqq \cdots \leqq a_nb_n$ are the eigenvalues of the matrix AB,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j'=1}^{N} F(\lambda_{Nj'}^{(jk)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F\left(\frac{f_j(x)}{f_k(x)}\right) dx. \tag{20}$$

The last result we need is that concerning the limiting joint distribution of quadratic forms in normal variates. Let

$$
\begin{aligned}
Q'_{jk} &= \frac{1}{N}\left(Q_{jk} - \frac{1}{2}\ln\frac{|\Sigma_{Nj}|}{|\Sigma_{Nk}|}\right) \\
&= \frac{1}{2N}\,\mathbf{x}_N{}'(\Sigma_{Nj}^{-1} - \Sigma_{Nk}^{-1})\,\mathbf{x}_N\,.
\end{aligned} \tag{21}
$$

We wish to find the limiting joint distribution of the set of random variables

$$
\bar{Q}_k{}' = \{Q'_{1k}\,, \cdots, Q'_{k-1,k}\,, Q'_{k+1,k}\,, \cdots, Q'_{mk}\}, \tag{22}
$$

for all $k = 1, \cdots, m$. The characteristic function of $\bar{Q}_k{}'$ is (cf. Grenander and Szegö (1958), p. 217, or Cramér (1946), p. 118)

$$
\phi_N(\bar{\alpha}_k) = \left| I_N - \frac{2i}{N}\sum_{\substack{j=1 \\ j\neq k}}^{m} \alpha_j(\Sigma_{Nk}\,\Sigma_{Nj}^{-1} - I_N)\right|^{-1/2}, \tag{23}
$$

where I_N is the identity matrix of order N, and

$$
\bar{\alpha}_k = \{\alpha_1\,, \cdots, \alpha_{k-1}\,, \alpha_{k+1}\,, \cdots, \alpha_m\}. \tag{24}
$$

Denoting the eigenvalues of the matrix

$$
\frac{1}{N}\sum_{\substack{j=1 \\ j\neq k}}^{m} \alpha_j(\Sigma_{Nk}\,\Sigma_{Nj}^{-1} - I_N)
$$

by $\lambda_{N1}(\bar{\alpha}_k), \cdots, \lambda_{NN}(\bar{\alpha}_k)$, Eq. (23) reduces to

$$
\ln \phi_N(\bar{\alpha}_k) = -\tfrac{1}{2}\sum_{j=1}^{N} \ln\,(1 - 2i\lambda_{Nj}(\bar{\alpha}_k)). \tag{25}
$$

Let

$$
2m_{jk} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{f_k(x)}{f_j(x)} - 1 - \ln\frac{f_k(x)}{f_j(x)}\right]dx, \tag{26}
$$

$$
4\sigma_{jk}^2 = \frac{2}{N}\,\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{f_k(x)}{f_j(x)} - 1\right)^2 dx, \qquad j, k = 1, \cdots, m. \tag{27}
$$

We consider the normalized stochastic variables

$$
Q''_{jk} = \sigma_{jk}^{-1}N^{-1}(Q_{jk} - m_{jk}). \tag{28}
$$

The characteristic function of

$$\bar{Q}_k'' = \{Q_{1k}'', \cdots, Q_{k-1,k}'', Q_{k+1,k}'', \cdots, Q_{mk}''\}, \tag{29}$$

is given by, cf. Eq. (25),

$$\ln \psi_N(\bar{\alpha}_k) = -\sum_{\substack{j,j'=1\\ j,j'\neq k}}^{m} \alpha_j \alpha_{j'} b_{jj'}^{(k)} + d_N + o(1), \tag{30}$$

since the asymptotic eigenvalue distribution of the sum of Toeplitz matrices is the same as the sum of the asymptotic eigenvalue distributions of the individual matrices, and where

$$| d_N | \leqq A \sum_{j=1}^{N} \left\{ \left[\sum_{\substack{j'=1\\ j'\neq k}}^{m} \left(\frac{\alpha_{j'}}{\sigma_{j'k}} \right)^2 \right]^{1/2} \left[\sum_{\substack{j'=1\\ j'\neq k}}^{m} (\lambda_{Nj}^{(kj')})^2 \right]^{1/2} \right\}^3, \tag{31}$$

$\lambda_{Nj}^{(kj')}$ are the eigenvalues of $(\Sigma_{Nk}\Sigma_{Nj'}^{-1} - I_N)/N$. According to Grenander and Szegö (1958), p. 219,

$$\sigma_{j'k} = O(N^{-1/2}), \qquad \text{all } j', k, \quad j' \neq k, \tag{32}$$

$$\lambda_{Nj}^{(kj')} = O(N^{-1}), \qquad \text{all } j, j', k, \quad j' \neq k, \tag{33}$$

so that

$$| d_N | \leqq AN^{-1/2} \cdot \sum_{\substack{j=1\\ j\neq k}}^{m} \alpha_j^2, \tag{34}$$

and we have

$$\lim_{N\to\infty} \psi_N(\bar{\alpha}_k) = \exp\left[-\sum_{\substack{j,j'=1\\ j,j'\neq k}}^{m} \alpha_j\, \alpha_{j'}\, b_{jj'}^{(k)} \right]. \tag{35}$$

Thus, the $(m - 1)$ normalized quadratic forms, $\bar{Q}_k''$, have a joint normal distribution with zero mean value matrix and covariance matrix $B_{m-1,k}$, which according to the regularity condition C is positive definite, cf. e.g. Anderson (1958), p. 36. It should be noted that there is no essential loss of generality in assuming $B_{m-1,k}$ is positive definite.

We obtain from Eq. (35) that

$$\lim_{N\to\infty} P_N'(k, R) = \lim_{N\to\infty} 1 - \int_{-m_{1k}/\sigma_{1k}}^{\infty} \cdots \int_{-m_{k-1,k}/\sigma_{k-1,k}}^{\infty} \cdot \int_{-m_{k+1,k}/\sigma_{k+1,k}}^{\infty} \cdots \int_{-m_{mk}/\sigma_{mk}}^{\infty} (2\pi)^{-(m-1)/2} \mid B_{m-1,k} \mid^{-1/2} \cdot \exp\left(-\mathbf{z}_{m-1}' B_{m-1,k} \mathbf{z}_{m-1}\right) dz_1 \cdots dz_{k-1}\, dz_{k+1} \cdots dz_m . \tag{36}$$

We also have that

$$-\frac{m_{jk}}{\sigma_{jk}} = -\frac{N^{1/2}}{2\sqrt{\pi}} \int_{-\pi}^{\pi} \left[\frac{f_k(x)}{f_j(x)} - 1 - \ln \frac{f_k(x)}{f_j(x)}\right] dx \Big/ \left[\int_{-\pi}^{\pi} \cdot \left(\frac{f_k(x)}{f_j(x)} - 1\right)^2 dx\right]^{1/2}. \tag{37}$$

Since $u - 1 \geqq \ln u$, $u \geqq 0$, the integral in the numerator of Eq. (37) is positive and we get

$$-\frac{m_{jk}}{\sigma_{jk}} \to -\infty, \qquad N \to \infty, \qquad \begin{matrix} j, k = 1, \cdots, m \\ j \neq k, \end{matrix} \tag{38}$$

so that

$$\lim_{N\to\infty} P_N'(k, R) = 0, \qquad k = 1, \cdots, m. \tag{39}$$

Thus, the probability of misclassifying an observation from C_k goes to zero, as $N \to \infty$, for all $k = 1, \cdots, m$, for the optimum Bayes procedure.

IV. ASYMPTOTICALLY UNCORRELATED VARIATES

If we are given a set of possibly complex-valued random variables $\{Y_{Nj}\}$, $j = 1, \cdots, N$, we say that these random variables are asymptotically uncorrelated if for each $\epsilon > 0$, there is a $N(\epsilon)$, such that

$$\mid E(Y_{Nj} - E(Y_{Nj}))(Y_{Nj'}^* - E(Y_{Nj'}^*)) \mid < \epsilon,$$

for all $N > N(\epsilon)$, and for each $j, j' = 1 \cdots, N, j \neq j'$. In particular, if U_N is a unitary matrix and if

$$\mathbf{y}_N = \begin{bmatrix} Y_{N1} \\ \vdots \\ Y_{NN} \end{bmatrix} = U_N \mathbf{x}_N , \tag{40}$$

then if $\{Y_{Nj}\}$ are asymptotically uncorrelated under C_k, $k = 1, \cdots, m$, the matrix U_N will be said to asymptotically diagonalize all Σ_{Nk}, and the process of obtaining the Y_{Nj}'s from the X_j's will be termed an asymptotic simultaneous diagonalization procedure (ASDP).

Let

$$U_N = \{N^{-1/2}\epsilon^{2\pi ijk/N}, j, k = 1, \cdots, N\}. \tag{41}$$

The matrix U_N is unitary since

$$\{U_N U_N'\}_{jk} = N^{-1} \sum_{j'=1}^{N} \epsilon^{2\pi i(j-k)j'/N} = \delta_{jk}. \tag{42}$$

We define

$$Y_{Nk_1} = N^{-1/2} \sum_{j_1=1}^{N} \epsilon^{2\pi i j_1 k_1/N} X_{j_1}, \qquad k_1 = 1, \cdots, N \tag{43}$$

and we will now show that this corresponds to an ASDP. We obtain

$$\begin{aligned} E_k(Y_{Nk_1} Y_{Nk_2}^*) &= N^{-1} \sum_{j_1,j_2=1}^{N} \epsilon^{(2\pi i/N)(j_1k_1-j_2k_2)} E_k(X_{j_1} X_{j_2}) \\ &= N^{-1} \sum_{j_1,j_2=1}^{N} \epsilon^{(2\pi i/N)(j_1k_1-j_2k_2)} \int_{-\pi}^{\pi} f_k(x)\epsilon^{-i(j_1-j_2)x} \frac{dx}{2\pi} \\ &= N^{-1} \int_{-\pi}^{\pi} f_k(x) \frac{dx}{2\pi} \sum_{j_1=1}^{N} \epsilon^{-ij_1(x-k_12\pi/N)} \sum_{j_2=1}^{N} \epsilon^{ij_2(x-k_22\pi/N)}. \end{aligned} \tag{44}$$

Now, we have

$$\sum_{j_1=1}^{N} \epsilon^{-ij_1(x-k_12\pi/N)} = \epsilon^{-i[(N+1)/2](x-k_12\pi/N)} D_N\left(x - k_1\frac{2\pi}{N}\right), \tag{45}$$

$$\sum_{j_2=1}^{N} \epsilon^{ij_2(x-k_22\pi/N)} = \epsilon^{i[(N+1)/2](x-k_22\pi/N)} D_N\left(x - k_2\frac{2\pi}{N}\right), \tag{46}$$

where

$$D_N(x) = \frac{\sin (N/2)x}{N \sin \frac{1}{2} x}. \tag{47}$$

Thus, using Eqs. (45), (46) in (44) we get

$$\begin{aligned} |E_k(Y_{Nk_1} Y_{Nk_2}^*)| = N \Bigg| \int_{-\pi}^{\pi} f_k(x) D_N\left(x - k_1\frac{2\pi}{N}\right) \\ \cdot D_N\left(x - k_2\frac{2\pi}{N}\right)\frac{dx}{2\pi}\Bigg|, \qquad 1 \leqq k_1, k_2 \leqq N. \end{aligned} \tag{48}$$

By using the series expansions given in Eqs. (45), (46) we obtain easily

$$N \int_{-\pi}^{\pi} D_N\left(x - k_1 \frac{2\pi}{N}\right) D_N\left(x - k_2 \frac{2\pi}{N}\right) \cdot \frac{dx}{2\pi}$$

$$\begin{aligned} &= 0, (k_1, k_2 = 1, \cdots, N, \\ &\qquad k_1 \neq k_2), \\ &= 1, (k_1 = k_2 = 1, \cdots, N). \end{aligned} \tag{49}$$

There is no loss of generality in considering only the cases $k_2 > k_1$, $k_1, k_2 \leqq N/2$. We introduce the numbers n tending to infinity with N in the following manner

$$n^4/N^3 = c, \tag{50}$$

where c is a generic constant independent of n, N. We have, for $|k_2 - k_1| \leqq 2n$, $k_1 \neq k_2$, using Eq. (49)

$$\begin{aligned} &\left| \int_{(k_1-n)2\pi/N}^{(k_2+n)2\pi/N} D_N\left(x - k_1 \frac{2\pi}{N}\right) D_N\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right| \\ &\leqq \int_{-\pi}^{(k_1-n)2\pi/N} + \int_{(k_2+n)2\pi/N}^{\pi} \left| D_N\left(x - k_1 \frac{2\pi}{N}\right) D_N\left(x - k_2 \frac{2\pi}{N}\right) \right| \frac{dx}{2\pi} \\ &\leqq \left(\frac{1}{\pi n}\right)^2 \left(\frac{2\pi - (k_2 - k_1 + 2n)2\pi/N}{2\pi}\right) \leqq \left(\frac{1}{\pi n}\right)^2, \end{aligned} \tag{51}$$

since

$$|D_N(x-\theta)| \leqq \frac{1}{N^{\frac{1}{2}}(n/N)2\pi} = \frac{1}{\pi n}, \qquad x \notin \left[\theta - \frac{n}{N} 2\pi, \theta + \frac{n}{N} 2\pi\right]. \tag{52}$$

Now, we can write

$$\begin{aligned} &\left| \int_{-\pi}^{\pi} N f_k(x) D_N\left(x - k_1 \frac{2\pi}{N}\right) D_N\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right| \\ &\quad \leqq \left| \int_{-\pi}^{(k_1-n)2\pi/N} \right| + \left| \int_{(k_1-n)2\pi/N}^{(k_2+n)2\pi/N} \right| + \left| \int_{(k_2+n)2\pi/N}^{\pi} \right| \\ &\quad \leqq \frac{2NM_k}{\pi^2 n^2} + N f_k \left[\frac{k_1 + k_2}{2} \frac{2\pi}{N}\right] \left| \int_{(k_1-n)2\pi/N}^{(k_2+n)2\pi/N} D_N\left(x - k_1 \frac{2\pi}{N}\right) \right. \\ &\qquad \left. \cdot D_N\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right| + N \left| \int_{(k_1-n)2\pi/N}^{(k_2+n)2\pi/N} \left(x - \frac{k_1 + k_2}{2} \frac{2\pi}{N}\right) \right. \\ &\qquad \left. \cdot f_k'(\hat{x}) D_N\left(x - k_1 \frac{2\pi}{N}\right) D_N\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right| \\ &\quad \leqq \frac{3NM_k}{\pi^2 n^2} + \left\{ \int_{(k_1-n)2\pi/N}^{(k_2+n)2\pi/N} \left| x - \frac{k_1 + k_2}{2} \frac{2\pi}{N} \right|^2 |f_k'(\hat{x})|^2 \right. \\ &\qquad \left. \cdot N D_N^2\left(x - k_1 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right\}^{1/2} \left\{ \int_{(k_1-n)2\pi/N}^{(k_2+n)2\pi/N} N D_N^2\left(x - k_2 \frac{2\pi}{N}\right) \right. \\ &\qquad \left. \cdot \frac{dx}{2\pi} \right\}^{1/2} \leqq \frac{3NM_k}{\pi^2 n^2} + \left(2n \frac{2\pi}{N}\right) |M_k'| = cN^{-1/4}. \end{aligned} \tag{53}$$

Therefore,

$$| E_k(Y_{Nk_1} Y^*_{Nk_2}) | \leqq cN^{-1/4}, \qquad | k_2 - k_1 | \leqq 2n. \tag{54}$$

If $| k_2 - k_1 | > 2n$ we obtain

$$\begin{aligned}
| E_k(Y_{Nk_1}\ Y^*_{Nk_2}) | &\leqq \left| \int_{-\pi}^{[(k_1+k_2)/2]2\pi/N} \right| + \left| \int_{[(k_1+k_2)/2]2\pi/N}^{\pi} \right. \\
&\quad \left. \cdot N f_k(x) D_N\left(x - k_1 \frac{2\pi}{N}\right) D_N\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right| \\
&\leqq \left\{ \int_{-\pi}^{[(k_1+k_2)/2]2\pi/N} f_k^2(x)\ N D_N^2\left(x - k_1 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right\}^{1/2} \\
&\quad \cdot \left\{ \int_{-\pi}^{[(k_1+k_2)/2]2\pi/N} N D_N^2\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right\}^{1/2} \qquad (55) \\
&\quad + \left\{ \int_{[(k_1+k_2)/2]2\pi/N}^{\pi} f_k^2(x)\ N D_N^2\left(x - k_1 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right\}^{1/2} \\
&\quad \cdot \left\{ \int_{[(k_1+k_2)/2]2\pi/N}^{\pi} N D_N^2\left(x - k_2 \frac{2\pi}{N}\right) \frac{dx}{2\pi} \right\}^{1/2} \\
&\leqq 2\left(M_k^2 N \frac{1}{\pi^2 n^2} \right)^{1/2} = cN^{-1/4}.
\end{aligned}$$

Therefore

$$| E_k(Y_{Nk_1} Y^*_{Nk_2}) | \leqq cN^{-1/4}, \qquad | k_1 - k_2 | > 2n. \tag{56}$$

Thus, $\{Y_{Nj}\}$ are asymptotically uncorrelated under C_k, $k = 1, \cdots, m$, and we have an ASDP based on the matrix U_N. The proof is independent of the choice of the constant c, although it may be different for each C_k. In addition, the proof does not require any assumption concerning the form of the probability density $p_k(\mathbf{x}_N)$. If $k_1 = k_2$, then it follows easily from the properties of Fejér's kernel, Grenander and Szegö (1958), p. 210,

$$E_k | Y_{Nk_1} |^2 \cong f_k\left(k_1 \frac{2\pi}{N}\right), \qquad k_1 = 1, \cdots, N. \tag{57}$$

V. CLASSIFICATION BASED ON AN ASDP

Since the $\{Y_{Nj}\}$ are uncorrelated under C_k, $k = 1, \cdots, m$, at least in an asymptotic sense, it might be supposed that these variates could be used in a classification procedure. That this is indeed possible will now be shown and the advantages of using such a procedure will also be pointed out.

We define

$$P_{jk} = \frac{1}{2}\left[\ln \frac{|F_{Nj}|}{|F_{Nk}|} + \mathbf{y}_N'(F_{Nj}^{-1} - F_{Nk}^{-1})\,\mathbf{y}_N\right], \qquad j, k = 1, \cdots, m, \quad (58)$$

where F_{Nk} is an Nth order diagonal matrix whose jth diagonal element is $f_k(j2\pi/N)$, $j = 1, \cdots, N$, $k = 1, \cdots, m$, and it is recalled that $\mathbf{y}_N$ is given in terms of $\mathbf{x}_N$ as in Eq. (43). The regions of classification are now $R_1' : P_{21} > 0, \cdots, P_{m1} > 0$; $R_2' : P_{12} > 0, P_{32} > 0, \cdots, P_{m2} > 0; \cdots;$ $R_m' : P_{1m} > 0, \cdots, P_{m-1,m} > 0$. This test, just as that based on the Q_{jk}'s, can be used even when $\mathbf{x}_N$ does not have a multivariate Gaussian distribution under C_k, $k = 1, \cdots, m$. This test is also quite simple to implement. The $\{Y_{Nj}\}$ are obtained from the $\{X_j\}$ by means of a simple linear transformation. The magnitudes of the $\{Y_{Nj}\}$ are obtained and weighted by constants which are determined by values of sampled power spectra. These quantities are summed, added to a bias term, which again is determined only by sampled power spectra, and the result yields the statistic on which the classification regions are based. This procedure is much simpler than the optimum Bayes test, since we no longer have to compute inverses and determinants of m Nth order matrices. However, the test based on the ASDP does lean very heavily on the assumption that the sampled power spectrum $f_k(x)$ is known for all $k = 1, \cdots, m$.

Let

$$P'_{jk} = \frac{1}{N}\left(P_{jk} - \frac{1}{2}\ln\frac{|F_{Nj}|}{|F_{Nk}|}\right) = \frac{1}{2N}\,\mathbf{x}_N'U_N'(F_{Nj}^{-1} - F_{Nk}^{-1})U_N\,\mathbf{x}_N, \quad (59)$$

and let

$$\bar{P}_k' = \{P'_{1k}, \cdots, P'_{k-1,k}, P'_{k+1,k}, \cdots, P'_{mk}\}, \qquad k = 1, \cdots, m. \quad (60)$$

The characteristic function of $\bar{P}_k'$, in the Gaussian case, is

$$\phi_N'(\bar{\alpha}_k) = \left| I_N - \frac{2i}{N}\sum_{\substack{j=1\\ j\neq k}}^{m} \alpha_j\, \Sigma_{Nk}\, U_N'(F_{Nj}^{-1} - F_{Nk}^{-1})U_N \right|^{-1/2}. \quad (61)$$

Now, it is easily seen, cf. Grenander and Szegö (1958), section 7.6, that the limiting eigenvalue distribution of the matrix

$$\frac{1}{N}\sum_{\substack{j=1\\ j\neq k}}^{m} \alpha_j\, \Sigma_{Nk}\, U_N'(F_{Nj}^{-1} - F_{Nk}^{-1})U_N$$

is the same as that of

$$\frac{1}{N} \sum_{\substack{j=1 \\ j \neq k}}^{m} \alpha_j (\Sigma_{Nk} \Sigma_{Nj}^{-1} - I_N).$$

This in turn implies that if

$$P''_{jk} = \sigma_{jk}^{-1} N^{-1} (P_{jk} - m_{jk}),$$

the limiting distribution of

$$\bar{P}_k'' = \{P''_{1k}, \cdots, P''_{k-1,k} P''_{k+1,k}, \cdots, P''_{mk}\}$$

is the same as that of $\bar{Q}_k''$. Thus, the $(m - 1)$ normalized quadratic forms, $\bar{P}_k''$, have a joint normal distribution with zero mean value matrix and covariance matrix $B_{m-1,k}$. In addition, the misclassification probabilities, of the classification system using the ASDP, all go to zero, as N approaches infinity.

VI. IMPLEMENTATION OF CLASSIFICATION SYSTEM USING ASDP

The implementation of the ASDP classification system is shown in Fig. 1. It is easily seen that the decision criterion used in this system corresponds to the classification regions $R_1', \cdots, R_m'$ defined previously. The system is seen to have considerable intuitive appeal from an engineering standpoint. If the observations are from C_k, then the system essentially first estimates $f_k(x)$, the sampled power spectrum under C_k, by means of 2 orthogonal sets of N narrow-band filters uniformly spaced from $2\pi/N$ to 2π. This estimate is then weighted by the reciprocal values of the sampled power spectrum, for each of the categories, and then combined with a bias term. The system then chooses C_k if the output of the kth channel is smaller than that of all the other channels. The system has been shown to have a performance which, at least asymptotically, compares favorably with the optimum Bayes system, under the Gaussian assumption. This is certainly enough justification for using the system in such applications. However, even in the nonGaussian case the system has considerable appeal due to the asymptotic simultaneous diagonalization property, which is independent of the Gaussian assumption. Thus, the system is generally applicable for classification problems dealing with stationary data.

It has, of course, been assumed that there is a learning phase during which $f_k(x)$ is estimated for each pattern. The output of the bank-of-filters shown in Fig. 1 can be used for this estimation procedure, cf. Grenander and Rosenblatt (1957). We mention briefly that if the f_k's are

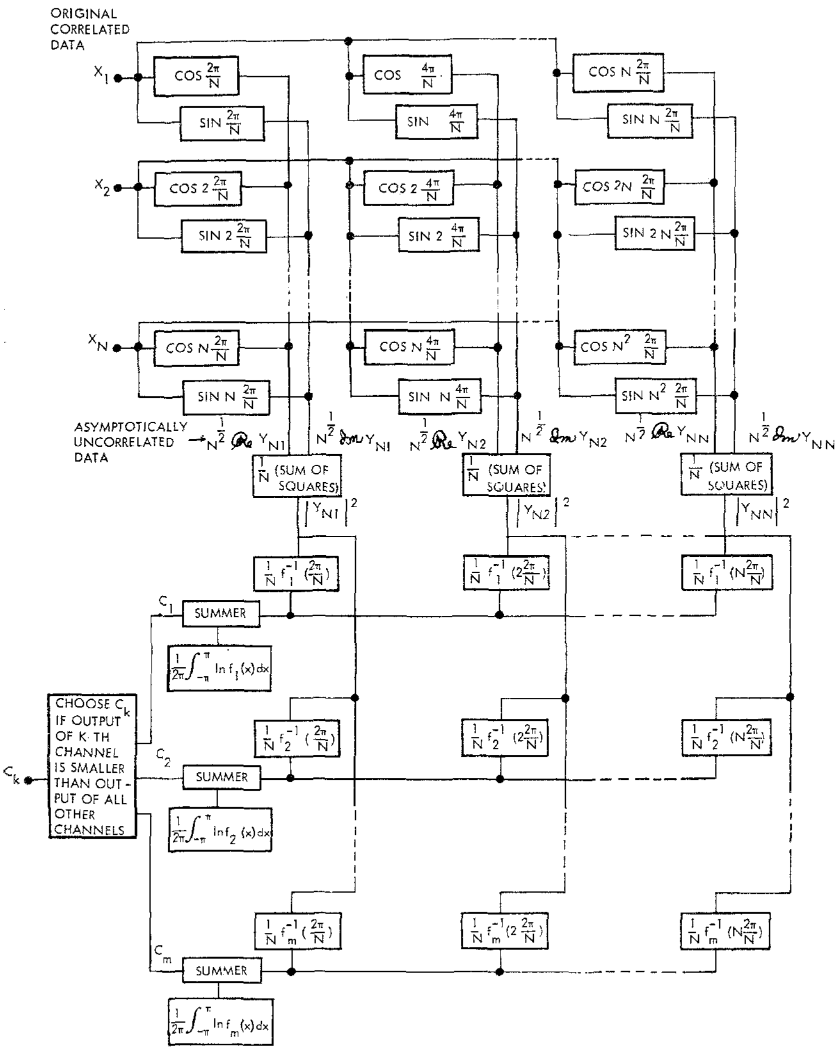

FIG. 1. Implementation of classification system based on asymptotic simultaneous diagonalization procedure.

allowed to vary slowly with time then the system of Fig. 1 can be made adaptive by using the output of the bank-of-filters, and the system's decision, to continually change the weights $f_k^{-1}(j2\pi/N)$, $j = 1, \cdots, N$, $k = 1, \cdots, m$.

The possible applications for the ASDP classification method are quite numerous. We mention, for example, radar or sonar, in which data is in sampled form, and it is desired to recognize several different targets, the returns from which can be considered as samples from a stationary random process.

It is interesting, and in fact desirable, to compare the present results with those of Price (1956) and Kailath (1960) for the detection of stochastic signals in noise. Since Kailath's work represents a generalization of Price's, only Kailath's results will be considered. The present work is closely related to Kailath's in the sense that both treat the problem of detecting one of m Gaussian random processes with known correlation functions. However, the underlying structures of the two problems are not the same so that the results obtained are quite different.

It is assumed by Kailath that the kth member of the set of m processes consists of a stochastic signal component which is different for each k and an additive independent noise component which is the same for each $k = 1, \cdots, m$. This type of assumption is not made in the present work, since no notion of additive noise enters into our discussion. In addition, Kailath assumes that the observed process can be processed continuously over a finite time observation interval. In the present work the observation consists of uniformly spaced samples, with spacing T, and the asymptotic results for N approaching infinity become valid when the total observation interval, NT, is quite large. It is the difference between these latter assumptions which makes Kailath's work quite different from ours.

The decision system obtained by Kailath is also quite different from the ASDP classification system. The optimum receiver of Kailath cross-correlates an estimate of the transmitted signal with the received signal and compares this value with a threshold, in order to determine which signal has actually been transmitted. This receiver structure is intuitively satisfying since it is a generalization of the usual matched filter receiver used for the detection of deterministic signals in additive Gaussian noise. In addition, Kailath's receiver structure has had some very successful applications in radar astronomy and scatter communication systems.

Although not mentioned explicitly by Kailath, it is possible for a certain type of singularity to occur in his problem in the sense that the misclassification probabilities can be made arbitrarily small over an arbitrarily small observation interval, cf. Slepian (1958). This singularity runs counter to engineering intuition, since it is known that no such result can be obtained in practice. However, it has been pointed out by several authors, cf. Wainstein and Zubakov (1962), and Root (1963), that by imposing some suitable regularity conditions on the spectra of the noise and signal processes it is possible to remove the singularity. These regularity conditions correspond to just the situations considered by Kailath, as well as Price, and explains why their results have been applied so successfully.

It is possible for a singularity to occur in the present work also, cf. Eq. (39). This type of singularity does not run counter to engineering intuition, however, since it states that the misclassification probabilities can be made arbitrarily small if we can employ an arbitrarily long observation interval. As a practical matter, it is, of course, impossible to obtain arbitrarily small error probabilities since the observation interval cannot be made arbitrarily large.

RECEIVED: March 30, 1964

REFERENCES

ANDERSON, T. W. (1958), "An Introduction to Multivariate Statistical Analysis." Wiley, New York.

ANDERSON, T. W., AND BAHADUR, R. R. (1962), Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.* **33**, 420–431.

CRAMÉR, H. (1946), "Mathematical Methods of Statistics." Princeton Univ. Press, Princeton, New Jersey.

GRENANDER, U., AND ROSENBLATT, M. (1957), "Statistical Analysis of Stationary Time Series." Wiley, New York.

GRENANDER, U., AND SZEGÖ, G. (1958), "Toeplitz Forms and Their Applications." Univ. California Press, Berkeley, California.

KAILATH, T. (1960), Correlation detection of signals perturbed by a random channel. *IRE Trans. Inform. Theory* **IT-6**, 361–366.

PERLIS, S. (1952), "Theory of Matrices." Addison-Wesley, Reading, Mass.

PRICE, R. (1956), Optimum detection of random signals in noise with applications to scatter-multipath communication, I. *IRE Trans. Inform. Theory* **IT-2**, 125–135.

RAGAZZINI, J. R., AND FRANKLIN, G. F. (1958), "Sampled-Data Control Systems." McGraw-Hill, New York.

ROOT, W. L. (1963), Singular Gaussian measures in detection theory. *In* "Time Series Analysis," M. Rosenblatt, ed. Wiley, New York.

SLEPIAN, D. (1958), Some comments on the detection of Gaussian signals in Gaussian noise. *IRE Trans. Inform. Theory* **IT-4**, 65–68.

WAINSTEIN, L. A., AND ZUBAKOV, V. D. (1962), "Extraction of Signals from Noise." Prentice-Hall, Englewood Cliffs, New Jersey.

WELCH, P. D., AND WIMPRESS, R. S. (1961), Two multivariate statistical computer programs and their application to the vowel recognition problem. *J. Acoust. Soc. Am.* **33**, 426–434.